Assessing the “Rothstein Falsification Test”
Does It Really Show Teacher Value-Added Models Are Biased?


Dan Goldhaber
The University of Washington

Duncan Chaplin
Mathematica Policy Research

October 31, 2011
The authors wish to acknowledge helpful comments on an earlier draft of this paper from Steve Glazerman, Mike Hansen, Eric Isenberg, Josh Kinsler, Cory Koedel, Arif Mamun, Alex Resch, Jesse Rothstein, Peter Schochet, Tim Sass, and Anastasia Semykina. We also would like to thank Stephanie Liddle and Roddy Theobald for their able assistance with this work. The views expressed in this paper are those of the authors and do not necessarily reflect the organizations with whom the authors are affiliated. Any errors are solely the authors’ responsibility.

The suggested citation for this working paper is:


© 2011 by Dan Goldhaber and Duncan Chaplin. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.
ABSTRACT

In a provocative and influential paper, Jesse Rothstein (2010) demonstrated that teachers assigned to students in the future are statistically significant predictors of past student achievement, a finding that obviously cannot be viewed as causal. Rothstein uses this finding to form the basis of a falsification test (the Rothstein falsification test), which is designed to indicate bias in estimates of teacher contributions to student learning that come from standard value-added models (VAMs).

Rothstein’s finding is significant because there is considerable interest in using VAM teacher effect estimates for policy purposes such as pay for performance or determining which teachers maintain their eligibility to teach after some specified period of time, such as when tenure is granted. If VAMs are shown to produce biased teacher effect estimates, it casts considerable doubt on the notion that they can be used to help inform high-stakes policy purposes, such as teacher tenure or pay.

In this paper we describe plausible conditions under which the Rothstein falsification test suggests bias when none exists, and provide results illustrating this point using Monte Carlo simulations. Our findings show that the Rothstein falsification test is not definitive in showing bias, which suggests a much more encouraging picture for those wishing to use VAM teacher effect estimates to help inform key policy decisions.
I. INTRODUCTION

In a provocative and influential paper, “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement,” Jesse Rothstein (2010) reports that VAMs used to estimate the contribution individual teachers make toward student achievement fail falsification tests, suggesting that VAM estimates are biased. Specifically, he shows that teachers assigned to students in the future have statistically significant predictive power in predicting past student achievement, a finding that obviously cannot be viewed as causal. Rather, the finding appears to signal that student-teacher sorting patterns in schools are not fully accounted for by the types of variables typically included in VAMs, implying a correlation between omitted variables affecting student achievement and teacher assignments. Rothstein presents this finding (his falsification test) as evidence of bias in VAM teacher effect estimates.

Rothstein’s falsification test has become a key method for academic papers to test the validity of VAM specifications. Koedel and Betts (2009), for instance, argue that there is little evidence of bias based on the Rothstein test when the VAM teacher effect estimates are based on teachers observed over multiple classrooms over time because the dynamic sorting is transitory. In their analysis of the impacts of teacher training, Harris and Sass (2010) report selecting school districts in which the Rothstein test does not falsify. Briggs and Domingue (2011) say that they use the Rothstein test to critique value-added results that were produced for the Los Angeles public schools and later publicly released.

---

1 In particular, sorting is dynamic in the sense that assignment to teachers is based on unobserved shocks to student achievement not accounted for by included covariates.
Rothstein’s finding has considerable relevance because there is great interest in using VAM teacher effect estimates for policy purposes such as pay for performance (Podgursky and Springer 2007; Eckert and Dabrowski 2010) or determining which teachers maintain their eligibility to teach after some specified period of time, such as when tenure is granted (Goldhaber and Hansen 2010a; Gordon et al. 2006; Hanushek 2009). If VAMs are shown to produce biased teacher effect estimates, it casts doubt upon the notion that they can be used for such high-stakes policy purposes. Indeed, this is how the Rothstein findings have been being interpreted. For instance, in an article published in Education Week (2009), Debra Viadero interprets the Rothstein paper to suggest that “‘value added’ methods for determining the effectiveness of classroom teachers are built on some shaky assumptions and may be misleading.” As Rothstein (2010) himself said, the “results indicate that policies based on these VAMs will reward or punish teachers who do not deserve it and fail to reward or punish teachers who do.” Indeed, in a recent congressional briefing, Rothstein cited his falsification test results and said that value-added is “not fair to… special needs teachers…[or] other specialists.”

In this paper, we show that the Rothstein falsification test can reject when there is no bias. More precisely, we describe plausible conditions under which the Rothstein test rejects the null hypothesis of no impacts of future teachers on lagged achievement even when there is no bias for estimated teacher effects. We verify these conditions theoretically and through a series of

---

2 Other researchers have come to somewhat different conclusions. Kane and Staiger (2008) find that some VAM specifications produce teacher effect estimates that are similar to those produced under experimental conditions. And while Koedel and Betts (2009) confirm Rothstein’s basic findings when it comes to single-year teacher effect estimates, they report finding no evidence of bias based on the Rothstein falsification test when the VAM teacher effect estimates are based on teachers observed in multiple classrooms over time.

3 Others have cited his work in similar ways (Hanushek and Rivkin 2010; Rothman 2010; Baker et al. 2010; and Kramer et al. 2011).

simulations. Our findings are important because, as noted above, the Rothstein test is shaping not only academic studies but also public perception about the efficacy of utilizing VAMs.

II. THE ROTHSTEIN FALSIFICATION TEST AND WHEN IT MIGHT BE WRONG

A. The Value-Added Model Formulation

There is a growing body of literature that examines the implications of using VAMs in an attempt to identify causal impacts of schooling inputs and contributions of individual teachers toward student learning (e.g., Ballou et al. 2004; Goldhaber and Hansen 2010; McCaffrey et al. 2004, 2009; Rothstein, 2009, 2010). Researchers typically assume their data can be modeled by some variant of the following equation:

\[ A_{ig} = \alpha + A_{i(g-1)} + \lambda + \sum \tau_{t,i,g} \beta + e_{ig} \tag{1} \]

where \( A_{ig} \) is the achievement of student \( i \) in grade \( g \),

\( \alpha \) is the intercept and also the value added of the omitted teacher,\(^5\)

\( \lambda \) is the impact of lagged achievement,

\( \tau_{t,i,g} \) is a dummy variable identifying if student \( i \) had teacher \( t \) in grade \( g \),

\( \beta \) is the impact of teacher \( t \) compared to the omitted teacher,\(^6\) and

\( e_{ig} \) is an error term that represents other factors that affect student learning.

If \( e_{ig} \) is uncorrelated with the other variables in equation 1, then the impact of teacher \( t \) can be estimated by regressing student achievement \( (A_{ig}) \) on prior student achievement \( (A_{i(g-1)}) \) and

\(^5\) The intercept equals the value of the outcome when all other variables are set to 0. If the achievement scores (current and baseline) are mean centered, the intercept equals the outcome for a student with an average baseline score who has the omitted teacher.

\(^6\) In the rest of this paper, we generally refer to \( \beta \) as the impact of teacher \( t \). This can be thought of as equivalent to saying that the results are normalized so that the impact of the omitted teacher is 0.
dummy variables identifying the teacher student i had in grade g (τt,i,g). Of course, this model makes a number of assumptions about the nature of student learning; see, for instance, Harris et al. (2010), Rothstein (2010), or Todd and Wolpin (2003) for more background on these assumptions.

Rothstein questions whether standard VAMs produce unbiased estimates of teacher effect on student learning. In particular, he describes a number of ways in which the processes governing the assignment of students to teachers may lead to erroneous conclusions about teacher effectiveness. Of particular concern is the possibility that students may be tracked into particular classrooms based on skills that are not accounted for by A_i,g-1.10

B. Bias in VAM Teacher Effect Estimates

If equation (1) is estimated using ordinary least squares, the estimated impacts of a teacher can be biased for a number of reasons. To derive a formula for this bias, we divide the error term (εig) into two components—ovig, that, if it exists, we assume to be correlated with at least some of

---

7 Like much of the value-added literature, Rothstein does not try to separate classroom and teacher effects.

8 This model is similar to the VAM2 model discussed by Rothstein (2010), which also allows the coefficient on lagged achievement to differ from equation 1 and excludes student-fixed effects. Value-added models often also include school and grade fixed effects and a vector of student and family background characteristics (for example, age, disability, English language status, free or reduced-price lunch status, race, ethnicity, whether a student has previously been retained in grade, and parental education). These factors could be incorporated into our models with changes in the substantive findings. Rothstein (2010) also discusses student-fixed effects and measurement error, two issues that we address below.

9 Parameter estimates can often be consistently estimated with large sample sizes but remain biased when sample sizes are small relative to the number of parameters being estimated. In this paper, we focus primarily on situations with large sample sizes. Kinsler (2011) investigates what happens to the Rothstein test when smaller sample sizes are used.

10 This could happen for a number of reasons. For example, principals might assign students to teachers based in part on unobserved factors that impact current achievement but are not captured by lagged achievement. Principals might do this either because they believe that certain teachers have a comparative advantage in educating certain students or because the principals are rewarding certain teachers with choice classroom assignments. Similar results could hold if parents lobby to have their children assigned to particular teachers.
the covariates included in the model even after controlling for the others, and \( eu_{ig} \), which is uncorrelated with all of the covariates in the model.

Thus,

\[
e_{ig} = \gamma \* ov_{ig} + eu_{ig}
\]

where \( \gamma \) is the coefficient on \( ov_{ig} \). This means that \( \gamma \) is also the coefficient that would be obtained on \( ov_{ig} \), were it added to equation 1.

It should be noted that by definition, if \( ov_{ig} \) exists it would cause bias for at least some coefficient estimates. However, as describe below, it can exist and still not bias estimated teacher effects.

The general formula for omitted variable bias for the effect of teacher \( t \) takes the form:

\[
Bias(\hat{\beta}_t) = E(\hat{\beta}_t) - \beta_t = \gamma \pi_e
\]

where \( \hat{\beta}_t \) = estimate of \( \beta_t \) and

\( \pi_e \) = coefficient on \( \pi_{t,i,g} \) from a regression of \( ov_{ig} \) on all right-hand side variables in equation 1 (except \( e_{ig} \)).

It can be shown that,

\[
\pi_e = \text{cov}(ov_{ig}^*, \pi_{v,i,g}^*) / \text{V}(\pi_{v,i,g}^*)
\]

(2) \(^{11}\)

where \( ov_{ig}^* \) and \( \pi_{v,i,g}^* \) are the conditional values of \( ov_{ig} \) and \( \pi_{v,i,g} \) that exist after controlling

\(^{11}\) Equation 2 holds because each coefficient estimate from any linear regression can be obtained by regressing the residualized outcome on the residualized version of the corresponding right-hand side variable in that equation. Each residualized variable is equal to the residual obtained after regressing the original right-hand side variable on all of the other right-hand side variables in the equation (Goldberger 1991).
for other variables in the model.\footnote{We use the abbreviation “cov” for “covariance” and “var” for “variance” in equations throughout the paper. We could also write, $\text{cov}_{\text{ig}} = \mathbf{x}_{\text{ig}} \pi + \text{cov}_{\text{igt}}$, where $\mathbf{x}_{\text{ig}}$ represents all the other covariates in the model and $\pi$ is a vector of coefficients on those variables. In our example, the vector $\mathbf{x}_{\text{ig}}$ includes lagged achievement and all other teacher variables in equation 1.}

This formulation is helpful because it shows that only the conditional value of $\text{ov}_{\text{ig}}$ matters for bias. Thus, $\text{ov}_{\text{ig}}$ might be correlated with lagged achievement and therefore bias that coefficient estimate, but still not cause bias in the estimated teacher effects if it is conditionally uncorrelated with $\tau_{\text{ig}}$, thus showing that omitted variables need not bias the estimated teacher effects.

In the standard VAMs discussed here, the conditional value of $\text{ov}_{\text{ig}}$ is based on a linear regression. Another way to state the condition of obtaining unbiased teacher effects in the presence of an omitted variable is thus: if the omitted variable is a linear function of lagged achievement and the other control variables in the equation, then it need not cause bias in the estimated teacher effects.

If the omitted variable is a nonlinear function of lagged achievement then it is more likely to cause bias. However, even in this case, it may still be possible to obtain unbiased estimates of the teacher effects if it is time-invariant (i.e. $\text{ov}_{\text{ig}} = \text{ov}_{t}$. It is possible, however, to address the issue of tracking through the inclusion of a student fixed effect, which captures time-invariant student or family background attributes that affect current achievement in ways not captured by previous achievement.\footnote{Studies differ on the efficacy of including student fixed effects—Harris and Sass (2010) argue for it, while Kane and Staiger (2008) argue against. Koedel and Betts (2009) find that student fixed effects are statistically insignificant in the models they estimate.} But, one of the primary contributions of Rothstein’s work is that he raises an additional concern about VAMs, postulating that student sorting into classrooms
is “dynamic” in the sense that an omitted student component of the error term may be both time-varying and cause bias. Specifically, he suggests that the omitted variables are negatively correlated over time. This could be due to compensating behavior whereby students who have a good year (along unobserved lines) receive fewer inputs in the following year than students who are otherwise similar. This dynamic form of tracking cannot be accounted for by the inclusion of a simple student fixed effect. In particular, Rothstein (2010) reports finding evidence suggesting negative correlations in errors over time in a model without student fixed effects. If important time-invariant omitted student factors exist, and implying the need for student fixed effects, we would expect to see a positive correlation across grade levels between the errors in a model estimated without student fixed effects. Given this, it seems unlikely that student fixed effects explain Rothstein’s finding of bias.

A. Rothstein’s Falsification Test

Rothstein’s falsification test relies on the notion that “future teachers cannot affect students’ past achievement,” so evidence of the statistical significance of future teachers in predicting past achievement suggests a misspecification of the VAM (Rothstein 2010). He bases his test, in part, on the fact that economists in many fields outside of education test for bias by estimating “impacts” of an intervention on past outcomes.

---

14 This could happen if family contributions to achievement are negatively correlated over time, conditional on past achievement.

15 One well-known paper that he cites is Ashenfelter (1978), who observes that earnings drop in an unusual way just before workers enter a training program, a phenomena now commonly known as the “Ashenfelter dip.” Because of this dip, Ashenfelter finds that estimated impacts of that training program depend critically on what baseline period is used. Similar patterns in test scores for certain teachers might also generate concern regarding VAMs.
We focus our paper on the version of the Rothstein test described in footnote 13 of his paper, where Rothstein states that the “exclusion restriction of random assignment conditional on $A_{ig-1}$ will be rejected if the grade-g classroom teachers predict $A_{ig-2}$ conditional on $A_{ig-1}$.”\footnote{It is important to note that the outcome in this equation is grade g-2 achievement. If one were to instead look at the impact of grade g teachers on grade g-1 achievement and control for grade g-2 achievement, then the nature of the test changes, although many of the same general properties hold. In particular, the test will often reject when there is no bias for estimated teacher effects, as we show in our simulations below.} Rejecting the null of no future teacher effects is then considered to be evidence of bias. In other words, it is assumed to imply a correlation between the independent variation in teacher assignments and the error in the achievement equation, a violation of the exclusion restriction of VAM specifications that student assignments to teachers are tantamount to random conditional on the covariates in the model.

A more precise formulation of the necessary exclusion restriction for unbiased teacher effects is that current grade teacher assignments are orthogonal to all residual determinants of the current score that are not accounted for by the exogenous variables included as controls in the model. This does not, however, mean that future grade teacher assignments need necessarily be orthogonal to the residual determinants of past scores in order for VAM models to produce unbiased estimates of current teacher effectiveness. In fact, tracking of future teachers based on lagged test scores would imply that past scores would be highly correlated with future grade teacher assignments. Rothstein’s falsification test does provide evidence of tracking and that tracking could cause VAM misspecification. But tracking based on prior achievement will also cause the Rothstein test to reject, even in the absence of any omitted variables that could cause bias.

\textbf{B. Aligning the Rothstein and Bias Tests}
The central connection between the Rothstein falsification test and a formal test of bias is through tracking of students, which can lead to bias in value-added estimates. However, tracking of students to future teachers (implied by the Rothstein test) need not mean that students are tracked to current teachers. For example, students can be tracked into future teacher classrooms based at least in part on past achievement, even if they were randomly assigned to current classrooms. Consequently, we assume that the Rothstein test relates most directly to the potential for bias in estimated impacts of future teachers on future achievement. Alternatively, by lagging the Rothstein test one period, and estimating impacts of current teachers on double-lagged achievement, it can be used to test for bias for current teachers. We lag the Rothstein test in part because it enables us to concisely compare the Rothstein and bias tests, as shown below. In addition, we note that if tracking systems are the same across grades, then it will not matter what grade we use when testing for bias or when doing the Rothstein test.

We utilize a simple VAM with only two teachers in each grade. The dummy for one is omitted from the regression and the current grade teacher assignment depends entirely on lagged achievement. Thus:

\[ A_{ig} = A_{i(g-1)}\lambda + \tau_{1,i,g} + e_{ig} \quad (3)^{17} \]

where \( \text{cov}(e_{ig},A_{i(g-1)})=\text{cov}(e_{ig},\tau_{1,i,g})=0. \)

We specify a flexible functional form for tracking for teacher 1 (the one with higher value added):

\[ \tau_{1,i,g} = T(A_{i(g-1)})^{18} \]

---

17 To simplify our presentation, we omit intercepts and measurement error from these equations. Adding them back does not change our substantive findings. Numerous methods exist to obtain unbiased estimates in the presence of such measurement error (Potamites et al. 2009; Meyer 1999; Fuller 1987).
The estimated impact of teacher 1 will be biased if

\[ \text{cov}(\tau_{1,i,g}, e_{ig})|A_{i(g-1)}) \neq 0 \]  

(4)

Rothstein’s falsification test is based on a regression of lagged achievement on current achievement and future teachers.\(^{19}\) He says that if future teachers are statistically significant in this regression, then that provides evidence of bias. We explain why the Rothstein test can reject the null hypothesis that future teachers have no effect on past student achievement, when there is no bias.

As noted earlier, to simplify our discussion we lag the Rothstein test one period so that instead of focusing on the relationship between future teachers and lagged achievement, the test is based on the relationship between current teachers and double-lagged achievement. This enables us to compare the conditions for bias and the Rothstein test using the same set of teachers. If the possible sources of bias are the same across grades, this simplification has no effect on our results.

The Rothstein test can be written as:

\[ A_{i(g-1)} = A_{ig} \lambda_1 + \tau_{1,i,(g+1)}R1_{(g+1)} + W_{i(g-1)} \]

where \( R1_{(g+1)} \) describes the relationship between future teachers and lagged achievement.

When we lag this one period, we get,

\[ A_{i(g-2)} = A_{i(g-1)} \lambda_2 + \tau_{1,i,g}R1_g + W_{i(g-2)} \]  

(5)\(^{20}\)

---

(continued)

\(^{18}\) \( T() \) is bounded between 0 and 1 and \( dT()/dA_{i(g-1)} > 0. \)

\(^{19}\) He uses another similar test in Table IV of his paper in which he adds in current teachers as an additional set of control variables. As shown in our simulations, this second test will also often reject when there is no bias.

\(^{20}\) As shown earlier, footnote 13 of Rothstein’s paper uses the same grade levels—g, g-1, and g-2.
The Rothstein test involves estimating whether or not $R_{1g}$ differs from 0. Alternatively, we can run the Rothstein test in two stages. In the first stage we omit current teachers.

$$A_{i(g-2)} = A_{i(g-1)}\lambda^3 + u_{i(g-2)}$$

(6)

Then, in the second stage, rather than estimating $R_{1g}$, we estimate the correlation between current teachers and the residual from equation 6. This is the numerator in $R_{1g}$.

$$\text{cov}(\tau_{1,i,g}, u_{i(g-2)}|A_{i(g-1)}) <> 0$$

(7)

For comparison, we repeat equation 4.

$$\text{cov}(\tau_{1,i,g}, e_{ig})|A_{i(g-1)}) <> 0$$

Formulated in this way, the Rothstein and bias tests (equations 7 and 4 respectively) appear quite similar but differ in a key respect—$e_{ig}$ is not the same as $u_{i(g-2)}$. One can, therefore, generate data in which there is no bias but the Rothstein test rejects (and vice versa). Specifically, this will occur if teacher assignments (i.e., tracking) depend on both lagged achievement and double-lagged achievement:

$$\tau_{1,i,g} = T(A_{i(g-1)}, A_{i(g-2)})$$

Of course the fact that one can come up with a data generating process that produces this particular result does not necessarily mean that this is a plausible process. But we argue that this process is actually quite plausible given that lagged test scores, as opposed to double-lagged test results, are often not available until sometime in the fall, too late for tracking decisions.

---

21 More precisely, $R_{1g} = \text{cov}(\tau_{1,i,g}, A_{i(g-2)}|A_{i(g-1)})/\text{var}(\tau_{1,i,g}|A_{i(g-1)}) = \text{cov}(\tau_{1,i,g}, u_{i(g-2)}|A_{i(g-1)})/\text{var}(\tau_{1,i,g}|A_{i(g-1)})$ since $u_{i(g-2)}$ is the residual that remains after regressing $A_{i(g-2)}$ on $A_{i(g-1)}$.

22 We provide examples of both of these situations in simulations below.

23 For instance, Mathematica does value-added work in various states and localities where teacher effect estimates are needed in a timely way to inform key policy decisions. In many of these locations, state test score data from the spring of one school year are often not available until the fall of the following school year, too late to impact tracking decisions for that year. See, for example, Potamites et al (2009) and Chaplin et al (2009).
at least some districts are likely to use double-lagged achievement rather than lagged achievement when making student assignment decisions. In addition, even when districts do not use double-lagged achievement explicitly, they may use other factors that are correlated with double-lagged achievement when making tracking decisions.

This need not cause bias if double-lagged achievement does not affect current achievement, as in equation 3. However, it will cause the Rothstein test to reject as long as \( A_{i(g-1)} \) is not identical to \( A_{i(g-2)} \). Under these conditions:

\[
\text{var}(A_{i(g-2)}|A_{i(g-1)}) > 0 \quad \text{and} \quad \text{cov}(\tau_{1,i,g}A_{i(g-2)}|A_{i(g-1)}) = \text{cov}(\tau_{1,i,g}, u_{i(g-2)}|A_{i(g-1)}) < 0.
\]

The example above illustrates only one reason that the Rothstein test might reject VAMs even when they produce unbiased teacher estimates. Here we break up the error term from equation 6 into three pieces to show other factors that might cause the Rothstein test to reject.

We start by repeating equation 6, the first stage of the Rothstein test.

\[
A_{i(g-2)} = A_{i(g-1)}^\lambda + u_{i(g-2)}
\]

Algebraic rearranging gives us,

\[
u_{i(g-2)} = A_{i(g-2)} - A_{i(g-1)}^\lambda
\]

Substituting out for \( A_{i(g-1)} \) gives us

\[
u_{i(g-2)} = A_{i(g-2)} - \lambda^3 (A_{i(g-2)}^\lambda + \tau_{1,i,(g-1)}\beta_{i(g-1)} + e_{i(g-1)})
\]

\[
= A_{i(g-2)}(1 - \lambda^3 \lambda) - \tau_{1,i,(g-1)}\lambda^3 \beta_{i(g-1)} - \lambda^3 e_{i(g-1)}
\]

Thus, the Rothstein test will reject if any of the following three conditions hold:

---

\[24\] This equation is similar to Rothstein’s equation 7 except that we put double-lagged achievement on the right hand side.
Exploring the Conditions

The first condition, as illustrated above, implies that if current teachers are selected in part based on double-lagged achievement, then the Rothstein test will reject. The second condition implies that if current teachers are selected at least in part based on past teachers, then the Rothstein test will reject. This is likely if some classrooms disproportionately consist of students who shared the same classroom in the previous year, perhaps because schools intentionally keep certain students together (or apart). We refer to this as “classroom tracking.”

The last condition might seem unlikely to matter since $e_{i(g-1)}$ is part of $A_{i(g-1)}$. Thus, one might assume that controlling for $A_{i(g-1)}$ would get rid of any possible correlation between $e_{i(g-1)}$ and $\tau_{i,i,g}$ and thus cause the Rothstein falsification test to not reject. However, this is not quite as clear as it first appears. Suppose, for example, that two groups of students have the same end of grade 4 test scores (grade $g-1$), but different grade 4 teachers—in this case there had to be factors, other than teacher quality, that explain the equality of end of grade 4 scores. It could be that the students with the less effective grade 4 teachers had higher grade 3 scores and thus similar grade 4 scores. If the grade 4 teachers are correlated with grade 5 teachers because of tracking, then grade 5 teachers can also be associated with grade 3 scores, within any group of

\[
\text{Cov}(\tau_{i,g}, A_{i(g-2)} | A_{i(g-1)}) \neq 0 \quad \text{C1}
\]
\[
\text{Cov}(\tau_{i,g}, \tau_{i,g-1} | A_{i(g-1)}) \neq 0 \quad \text{C2}
\]
\[
\text{Cov}(\tau_{i,g}, e_{i(g-1)} | A_{i(g-1)}) \neq 0 \quad \text{C3}
\]

There are alternative tests for bias caused by these types of tracking. For example, for the first condition, one can include double-lagged achievement in the VAM model and test to see if the estimated coefficient estimates on current teachers change compared to a model without that variable (Rothstein 2009). Similarly, for the second condition, one can add lagged teachers to a standard VAM. Results of this later test are likely to be very imprecise for many teachers, especially in smaller schools, if most of their students come from a single lagged teacher.
students that have the same grade 4 scores (i.e., after controlling for those scores). This means that grade 5 teachers can be statistically significant in the Rothstein test regression, which involves regressing grade 3 scores on grade 4 scores and grade 5 teachers. Thus, it may be possible for the Rothstein test to reject, even in the absence of conditions 1 and 2. We show this below with a simulation and discuss it further in Appendix A.

This third condition is important because Rothstein (2010) expresses particular interest in the distribution of error terms. In particular, he acknowledges that evidence that future teachers are statistically significant predictors of past achievement is not itself proof of bias for current teachers. He goes on to say that tracking accompanied by negative correlation in the errors across grades (i.e., \( \text{cov}(e_{ig}, e_{(i,g-1)}) < 0 \)) “strongly suggests” bias for current grade teachers. More precisely he says,

“A correlation between treatment and some pre-assignment variable X need not indicate bias in the estimated treatment effect if X is uncorrelated with the outcome variable of interest. But outcomes are typically correlated within individuals over time, so an association between treatment and the lagged outcome strongly suggests that the treatment is not exogenous with respect to post-treatment outcomes (Rothstein 2010).”

We agree that if the errors are negatively correlated and there is tracking then the teacher effects will probably be biased (see Appendix B) and the Rothstein test will reject based on condition 3 (see Appendix A). However, the Rothstein test will reject based on conditions 1, 2, or 3 even without negatively correlated errors. Thus, one cannot use the Rothstein test to identify bias caused by negatively correlated errors.

---

26 A negative correlation could occur if students who were above average based on their prior skills forget more of what they learned in the past than other students, and/or those who were below average learn more than other students from their current classmates and/or teacher.
That said, the Rothstein test does work if one assumes that error terms are negatively correlated but remains agnostic about the presence of tracking. Under these conditions (as shown in our simulations below), the Rothstein test will only reject if there is tracking and there will only be bias if there is tracking. Consequently one can use the Rothstein test to identify bias caused by tracking based on lagged achievement under these conditions.

While the Rothstein test cannot be used to identify bias caused by negatively correlated errors, Rothstein presents other evidence that suggests negatively correlated errors. In general, this would suggest bias, as shown in our simulations. The bias could be quite small (Rothstein 2009; Kinsler 2011; and Appendix C of this paper). In addition, one might question Rothstein’s evidence on the magnitude of negatively correlated errors. First, as shown in Appendix D, it is difficult to distinguish between negatively correlated errors in a value-added model and measurement error. Second, the Rothstein evidence is based on the correlation in growth rates across grades and not on a correlation estimate based on the results of one of his value-added models. Third, there is uncertainty regarding how much test measurement error exists. For example, different methods of estimating the measurement error usually come up with different estimates. This means that it is difficult to know how much of the negative correlations he observes could be due to measurement error. Last, but not least, Rothstein finds almost no

---

27 For example, allowing the coefficient on lagged test scores ($\lambda$) to be less than one could matter if prior learning decays over time. This is because decay in impacts of the factors captured in the error term might cause negatively correlated observed errors. However, allowing $\lambda$ to be less than one implies decay in the impacts of all past inputs that impact lagged achievement. This includes double lagged achievement, lagged teacher impacts, and the lagged error. If these different components decay at different rates then $\lambda$ might not capture the decay appropriately. A priori it is not clear whether the error would decay faster or slower than the other components. Consequently, it is not clear whether differential decay would cause a positive or negative correlation in the resulting errors.

28 These methods include test-retest, which involves comparing results across different administrations of the same test, and internal consistency, which involves comparing results across different items within the same test.
correlation in growth rates separated by a year so that they share no common test. This is exactly the pattern one would expect to see for the correlation between errors across years if it were caused by measurement error.

III. SIMULATION RESULTS

We perform a number of simulations that illustrate the findings reported in the prior section. For consistency with Rothstein (2010), we simulate data for grades 3 through 5. We conduct Rothstein falsification tests (described by equation 5 above) and tests for bias for the estimated grade 5 teacher effects. The errors have standard deviations of 0.2 and the coefficient on lagged achievement is set to 0.98. Teacher effects in grades 3 and 4 are set to either 0.1 or 0 (depending on the model). Students are tracked based on indicator variables—those with above average values on the indicator are given teacher 1 and those with below average values are with the omitted teacher. The tracking indicator is the sum of the components used for tracking, which are indicated in Table 1 (below) for each model. The factors used for tracking include previous achievement, double-lagged achievement, previous teacher dummy, and a random component that has a standard deviation of 0.1. In some models, the achievement error terms have a negative correlation of around -0.24.

---

29 We start with normally distributed achievement in grade 2 and then add in teacher effects and errors for achievement in grades 3, 4, and 5.

30 The grade 5 teachers can be thought of as future teachers in grade 4 using the regular Rothstein test or current teachers using our revised test (lagging the Rothstein test one period). As noted earlier, if tracking systems are stable across grades, then the choice of grade levels will not matter.

31 The choice of the coefficient on lagged achievement does not impact our substantive findings as long as it is not 0. We set this coefficient to 0.98 in part to generate achievement data with standard deviations close to 1.

32 Rothstein simulates data in Appendix C of his paper using a negative correlation of -0.25.
In some models we include an omitted variable that impacts tracking decisions and current test scores (by adding it to the tracking indicator and achievement, respectively), but is not correlated with lagged achievement. It is given a standard deviation of 0.2.

For each model we test to see if the model is rejected using the Rothstein falsification test and also if the teacher impact estimate in grade 5 is biased. We simulate data with either 1 million or 100,000 students, depending on the model.

Achievement levels in grades 3, 4, and 5 generated using this process have standard deviations ranging between 1 and 1.1 (depending on the model and grade). This means that all other factors are in approximate effect size units.

In Table 1 we present five sets of results. The first set of columns (under “Results by Condition”) demonstrates how the Rothstein test performs based on the three conditions discussed above. The second set of columns (under “Normal Baseline Scores”) covers what happens if grade 4 achievement is normally distributed, which is important because it helps to explain why the Rothstein test rejects based on the third condition described above. The third set of columns (under “Negatively Correlated Errors”) describes conditions under which the Rothstein test works. The fourth set of columns (under “Failing to Falsify”) presents cases where the Rothstein test fails to falsify when it should. The last set of columns (under “Rejecting RA”) shows how the Rothstein test can reject in spite of random assignment (RA).

---

33 This is a point he acknowledges is possible in his paper.
Table 1: Simulation Results for Rothstein Falsification Test and Bias Test by Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Result</th>
<th>Normal Baseline Scores</th>
<th>Negatively Correlated Errors</th>
<th>Failing to Falsify</th>
<th>Rejecting RA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Results by Condition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C1 C2 C3</td>
<td>C1 C2 C3</td>
<td>M1 M2 M3</td>
<td>M1 M2</td>
</tr>
<tr>
<td>G5 Teacher 1 Tracked on</td>
<td>A₄, A₃ A₄τ₄ A₄</td>
<td>A₄ A₄τ₄ A₄</td>
<td>Rand A₄ A₄ Rand A₄ τ₄ τ₄</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4 Teacher 1 Tracked on</td>
<td>A₃ A₃ A₃</td>
<td>A₃ A₃ A₃</td>
<td>Rand A₃ A₃ Rand A₃ A₃τ₃</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4 Teacher 1 Effect</td>
<td>0.1</td>
<td>0.1 0.1</td>
<td>0.1 0.1</td>
<td>0 0 0</td>
<td>0.1 0.1</td>
</tr>
<tr>
<td>G3 Teacher 1 Effect</td>
<td>0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0.1 0.1</td>
</tr>
<tr>
<td>Cor(ₑᵣ₉,ₑᵣ₉₋₁)= Cor(ₑᵣ₉₋₁,ₑᵣ₉₋₂)</td>
<td>0</td>
<td>0 0 0</td>
<td>-0.24 -0.24 -0.24 0 -0.25 0</td>
<td>0 0</td>
<td></td>
</tr>
<tr>
<td>Var(ₒᵥᵣᵣ)</td>
<td>0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0.2 0.2 0 0</td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>100K 100K 100K</td>
<td>100K 100K 1 Mil.</td>
<td>100K 1 Mil. 100K 100K 1 Mil.</td>
<td>1 Mil. 1 Mil. 1 Mil.</td>
<td></td>
</tr>
<tr>
<td>Rothstein T-Test</td>
<td>9</td>
<td>31 -21</td>
<td>41 78 0.96</td>
<td>-0.49 -0.93 -1.06 -0.93 -62 25</td>
<td></td>
</tr>
<tr>
<td>Bias T-Test</td>
<td>0.46</td>
<td>0.97 -0.10</td>
<td>-0.74 -0.82 0.74 1.34 9.67 -0.39 37 54 0.43 0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias Point Estimate</td>
<td>0.0010</td>
<td>0.0023 -0.0002</td>
<td>-0.0015 -0.0017 0.0005 0.0017 0.0065 -0.0008 0.0670 0.1500 0.0002 0.0003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Shading indicates parameter(s) that changed from previous model. In all models the grade 5 teacher impact is 0 so the bias point estimate is the same as the point estimate for the grade 5 teacher. 100K = 100,000 and 1 Mil. = 1 million. The random errors that help determine tracking for grades 4 and 5 have standard deviations of 0.1. Rand means that this random error was the only factor used for tracking. The standard deviations of the errors are all 0.2. The grade 2 achievement has a variance of 1. The probability of having teacher 1 in each grade is 0.5. We use 1 million cases for some models to obtain more precise estimates and/or to ensure that our finding of unbiased results was not likely caused by small sample size. The variables eᵣₑᵣ₋₁ and eᵣ₋₂ are error terms in the achievement equations. The omitted variable, oᵥᵣᵣ, is a component of eᵣ that affects both achievement and tracking in grade 5, but that is uncorrelated with all previous grade variables.
We start by showing cases where the Rothstein rejects but there is no bias, relating it to the three conditions discussed above. For the first two conditions, the reason for the Rothstein test to reject seems fairly clear—current teachers are selected in part based on double-lagged achievement and/or lagged teachers, both of which are components of the error term from the first stage of the Rothstein test. Hence, current teachers are correlated with that error. In the third model, however, neither condition is present and yet the Rothstein test still rejects. Here the false rejection it is caused by the fact that past teacher effects can create a relationship between current teachers and double-lagged achievement conditional on lagged achievement, as discussed earlier. The results in columns four through nine of Table 1 help to illustrate this and Appendix A provides another explanation for why this is possible.

In the columns under “Normal Baseline Scores,” we generate data with normally distributed baseline (grade 4) scores by setting grades 3 and 4 teacher effects to 0. This works because all other components of the baseline test are normally distributed (achievement in grade 2 and the error terms). Only the teacher variables are not normal (discrete in our case). Thus, if the teacher effects are set to zero, the resulting achievement variables will be normally distributed.

As illustrated in the first two columns under “Normal Baseline Scores” in Table 1 the Rothstein test still rejects if tracking is based on either grade 3 achievement or grade 4 teachers.

---

34 In Table 1 we present results using the Rothstein test described in equation 5 above. We also ran two variations of the Rothstein test. The first involved adding in grade 4 teachers to equation 5. We refer to this as the Rothstein 2 test. He also used this in his paper. The second variation involved switching grade 3 and 4 achievement in equation 5. Thus, for this variation, we regress grade 4 achievement on grade 3 achievement and grade 5 teachers. We call this the Rothstein 3 test. It was not used in his paper. For the first three columns of Table 1, the results for these alternative tests were similar to those presented in the table in the sense that the coefficient on the grade 5 teacher was statistically significant with t-stats ranging from 16 to 103 in absolute magnitude.

35 The alternative versions of the Rothstein test rejected models C1, C2, and C3 under “Normal Baseline Scores.” The smallest t-stat on the grade 5 teacher was 3 for the Rothstein 3 test, C2. The other t-stats ranged from 22 to 255 in absolute magnitude. This means that the results differ for C3 since the main Rothstein test (presented in our Table 1) did not reject that specification.
However, if neither of those conditions hold, then the Rothstein test no longer rejects, as can be seen in the third column under “Normal Baseline Scores.”\textsuperscript{36} This is important because it highlights the fact that the Rothstein test is rejecting in part because of the existence of grade 3 and 4 teacher effects. Those effects cause grade 4 achievement to be non-normally distributed so that it can no longer completely adjust for the relationship between the error from our equation 6 (stage 1 of the Rothstein test) and the current teacher. This, in turn, caused the Rothstein test to reject based on condition 3 in column 3 (without normal baseline scores). The important role of normally distributed baseline scores is discussed further in Appendix B.

In the columns under “Negatively Correlated Errors,” we present the conditions under which the Rothstein test does identify bias. In particular, if the error terms are negatively correlated but there is no tracking, the Rothstein test does not reject and there is no bias (column M1). If there is tracking, then the Rothstein test rejects and there is bias (column M2).\textsuperscript{37} The bias term has an effect size of less than 0.01.\textsuperscript{38} In column M3, we show that if the baseline scores had been normally distributed, then the Rothstein test would not reject and there would be no bias. This is because under those conditions the baseline test can control for the negatively correlated errors, as discussed in Appendix B. Thus, as with the Rothstein test results, the bias results are also affected by the normality of the baseline score. To the extent that the baseline scores are not

\textsuperscript{36} We used 1 million observations in this simulation to help ensure that the results were not caused by a small sample size.

\textsuperscript{37} A similar result holds with measurement error if tracking is done based on observed achievement (including the measurement error). When baseline scores are normally distributed, there is no bias and the Rothstein test does not reject. With non-normal baseline scores, his test does reject and there is bias.

\textsuperscript{38} We used 1 million cases in this simulation instead of the standard 100,000 so that we could get a more precise estimate of the bias term. Doubling the negative correlation between the errors (to -0.50) roughly doubles the magnitude of the bias (to around 0.012). Doubling the magnitude of the grade 4 teacher effect (to 0.20) and keeping the negative correlation at -0.50, doubles the magnitude of the bias again (to around 0.025).
normally distributed, some bias can result. However, in Appendix C we present simulations suggesting that, despite bias, there still exists a high correlation between value-added estimates and true teacher effectiveness in a model with negatively correlated errors.

If we knew that the error terms were negatively correlated but were unsure if there is tracking, then the Rothstein test clearly would yield the correct finding. However, given that tracking in schools is quite likely, the assumption of negatively correlated errors is tantamount to assuming bias. More generally, and as discussed earlier, we cannot use the Rothstein test to check for bias caused by negatively correlated errors because we cannot tell if it is rejecting because of those or because of one of the other conditions.

In the columns under “Failing to Falsify” in Table 1, we present models in which the Rothstein test fails to falsify, but the results are biased. We get this by creating an omitted variable that affects both grade 5 achievement and the selection of the grade 5 teachers, but which is uncorrelated with lagged achievement and lagged teachers. Thus, it does not cause the Rothstein test to reject. In the second column, we estimate a model with negatively correlated errors and the Rothstein test still fails to reject—although we note that this is because the baseline test is normally distributed. If the baseline test were not normally distributed, then the Rothstein test would have rejected.

---

39 For the results under “Negatively Correlated Errors,” the Rothstein 2 and 3 tests produced similar results to those for the first model (t-stats below 1 in absolute value) and for the second model (t-stats over 191 in absolute value). For the third model, the Rothstein 2 and 3 tests rejected with t-stats of 61 and above in absolute value and thus differed from the nonsignificant result for the Rothstein test based on his equation 11.

40 The alternative versions of the Rothstein test produced the same result for the first column under “Failing to Falsify” (nonsignificant results with t-stats close to 1). For the second column, the two alternative versions of the test both rejected with t-stats of -48 and 78 respectively.
Random assignment of individual students to teachers is a clear case where VAMs yield unbiased estimated teacher effects, and the Rothstein test appropriately fails to falsify. Interestingly, however, random assignment of groups of students to teachers can cause the Rothstein test to falsify if students were tracked into those groups. This is shown in the last two columns of Table 1, which report findings when students are tracked into classrooms based on their previous classrooms alone, but where teacher assignment to classrooms is random.\footnote{The Rothstein 2 test, which controls for grade 4 teachers, did not reject either model in the last two columns of Table 1. This is not surprising since tracking is based only on grade 4 teachers in this model. The Rothstein 3 test did reject both models with t-stats of 36 and above.} The same arguments for Rothstein rejecting hold as for the previous models. Indeed, the first three columns of Table 1 are relevant because they simply describe how students are tracked into grade 5 classrooms, but not how grade 5 teachers are assigned to those classrooms. Thus, they would hold if teachers were randomly assigned.\footnote{Kinsler (2011) finds a similar result for the Rothstein test of a “VAM1” model in which the outcome for the VAM is test score growth and there is no control for lagged achievement.} This is particularly interesting because the Gates Foundation is currently conducting a large random assignment study of teacher evaluation systems using this type of random assignment (MET Project 2010). Our results suggest that the Rothstein test would reject teacher effectiveness estimates produced using the random assignment methods used by the Gates study if students are tracked into classrooms before teachers are randomly assigned.
IV. CONCLUSION

As we noted in the outset of this paper, Rothstein’s critique of value-added methods used to estimate teacher effectiveness has been cited by both the research and policymaking communities as a reason to doubt the wisdom of using VAMs for high stakes purposes. If, as Rothstein’s falsification test suggests, VAMs produce biased estimates of individual teacher effectiveness, this policy direction may indeed be misguided. The findings we present here, however, call into question whether the Rothstein falsification approach provides accurate guidance regarding the bias of teacher effect estimates.

Rothstein’s test certainly provides evidence that students are tracked, but tracking of the type indicated by the Rothstein test need not cause bias in estimated teacher effects. More precisely, we have identified two conditions in which this is true. The first is when tracking is based on double-lagged achievement of students assigned to a teacher’s classrooms, a condition that may be common when lagged achievement data are not ready in time for the tracking decisions. A similar result holds if students are tracked in part on the classroom they were in during the previous school year. In both situations, the Rothstein test can falsify when it should not.

We would argue that Rothstein’s 2010 paper raised important concerns about the ability of VAMs to produce unbiased estimates of teacher effectiveness but the Rothstein test itself does not definitively falsify them. This implies that more work should be done to investigate factors that might cause bias in value-added models. One way to approach this topic is to look at the various factors affecting student sorting into classrooms other than those typically included as
controls in value-added models. Work like this has been done by numerous authors both quantitatively (Jacob and Lefgren 2007) and qualitatively (Kraemer et al. 2011). When doing such work, however, researchers may want to keep in mind our results which suggest that some variables often omitted from VAMs (for example double-lagged achievement and lagged teachers) may affect the selection of current teachers and yet not cause much bias.

From a policy perspective, the important questions may not be whether there is any bias, but its magnitude. It is quite likely that teacher effectiveness estimates generated from VAMs are biased to some degree, but as shown in Rothstein (2009), Kinsler (2011), and in our Appendix C, the magnitude of bias may be relatively inconsequential. Decisions about using VAM should consider how this bias compares to potential information that value-added models can provide about teacher effectiveness over, or in addition to, other means of assessment.\(^\text{44}\)

---

\(^{43}\) Ashenfelter (1978) makes a similar point in his paper that looks at similar issues outside of value-added models.

\(^{44}\) Value added estimates may also be very imprecise (Schochet and Chiang, 2010). Other measures of teacher effectiveness may also be imprecise so the policy focus should probably be on how best to obtain more precise estimates. This may involve some combination of VAM and non-VAM measure.
REFERENCES


APPENDIX A

ROTHSTEIN TEST CAN REJECT BASED ON CONDITION 3

In the main body of this paper, we provided one explanation for why the Rothstein test can reject based on Condition 3. In this appendix, we provide an alternative explanation. As stated in the main body of this paper, this condition means that the Rothstein falsification test will reject even in the absence of negatively correlated errors, tracking based on double-lagged achievement, or tracking based on lagged teachers. More precisely, Condition 3 is that the grade 5 teacher is correlated with the grade 4 error term, conditional on grade 4 achievement:

$$\text{cov}(\tau_{5,t}, g, e_{t(g-1)} | A_{t(g-1)}) \neq 0$$ (C.3)

If all three of the variables in Condition 3 were jointly normally distributed, then this would not be possible. This is because if the error term and lagged achievement are jointly normal, then the expected value of the error would be a linear function of lagged achievement and the conditional error would be uncorrelated with any function of lagged achievement (linear or nonlinear), as discussed in Appendix B.

If lagged teachers impact lagged achievement, then it is likely that lagged achievement is not normally distributed. Consequently, the expected value of the lagged error can be a nonlinear function of lagged achievement, as is the current teacher. Since both are nonlinear functions of lagged achievement, they can remain correlated even after conditioning on lagged achievement in a linear regression. This could, in turn, cause the Rothstein test to reject, as suggested by our simulations.
APPENDIX B

NEGATIVELY CORRELATED ERRORS CAN CAUSE BIAS

In this appendix we show that negatively correlated errors need not cause bias if lagged test scores are normally distributed. We then show that since baseline test scores are not likely to be completely normally distributed, some bias is likely. To make the first point, we first show that an omitted variable that is a linear function of the lagged test score need not cause bias, a point that is well known. This is important because it means that the negative correlation in errors across grades assumed by Rothstein need not cause bias on its own. Second, we show that negatively correlated errors and a non-normal distribution of lagged test scores caused by lagged teacher effects could cause bias in estimated effects for current teachers. This supports Rothstein’s model because if we assume negatively correlated errors (and lagged teacher effects), then evidence of tracking based on lagged test scores (which the Rothstein test implies) suggests bias for current teachers. In Appendix C, we show that the bias can be quite small.

No Bias if Omitted Variables are Linear Functions of Lagged Test Scores

To investigate these issues, we consider a model in which the true error (not the measurement error) is negatively correlated with the lagged error, as Rothstein posits. This could happen if students who had an above average error term in the previous period forget more than students with an average error term in direct proportion to how far above average they were in the previous period. Similarly, those who had a below average error term in the previous period learn more than students with an average error term (perhaps from other students or their teacher) again, in direct proportion to how far they were below average in the previous period.\textsuperscript{45}

\textsuperscript{45} This is sometimes described as “regression to the mean,” although that phrase is sometimes used to describe situations in which the true errors are uncorrelated across grades.
Mathematically this can be written as:

\[ e_{ig} = \rho^* e_{iug(g-1)} + e_{iug} \]

where \( \text{cov}(e_{iug}, e_{iug(g-1)}) = 0 \) and \( \rho < 0 \).

This implies that \( \text{cov}(e_{ig}, e_{i(g-1)}) = \rho^* \text{var}(e_{iug(g-1)}) < 0. \) (B.1)

As in our simulations, we assume that lagged achievement depends only on lagged teachers, an error term, and a random starting point \( (A_{i(g-2)}) \) that is normally distributed.\(^{47}\) Thus:

\[ A_{ig} = A_{i(g-1)} + \lambda + \tau_{i,g} + e_{ig} \]
\[ A_{i(g-1)} = A_{i(g-2)} + \tau_{i,g-1} \beta_{i(g-1)} + e_{i(g-1)} \]

To generate normally distributed baseline scores (and consistent estimated teacher effects) we assume that lagged teachers have no impact on lagged achievement levels. Thus,

\[ A_{i(g-1)} = A_{i(g-2)} + e_{i(g-1)} \]

where \( A_{i(g-2)} \), \( e_{i(g-1)} \) and \( e_{ig} \) are jointly normal and uncorrelated with each other by assumption. This implies that \( A_{i(g-1)} \) and \( e_{ig} \) are also jointly normal and linearly related because a linear function of two jointly normally distributed variables is also jointly normal and linearly associated with each of those variables (Goldberger 1991). Indeed, all four variables are jointly normal and linearly related.

\(^{46}\) In the following five lines, we derive equation (B.1).

\[ \text{cov}(e_{ig}, e_{i(g-1)}) \]
\[ = \text{cov}(\rho^* e_{iug(g-1)} + e_{iug}, \rho^* e_{iug(g-2)} + e_{iug(g-1)}) \]
\[ = \rho^2 \text{cov}(e_{iug(g-1)}, e_{iug(g-2)}) + \rho^2 \text{cov}(e_{iug(g-1)}, e_{iug(g-1)}) + \rho^2 \text{cov}(e_{iug}, e_{iug(g-2)}) + \text{cov}(e_{iug}, e_{iug(g-1)}) \]
\[ = \rho^2 \cdot 0 + \rho^2 \cdot \text{var}(e_{iug(g-1)}) + \rho^2 \cdot 0 \]
\[ = \rho^2 \cdot \text{var}(e_{iug(g-1)}) \leq 0. \]

\(^{47}\) This can be justified if we think of two grade levels in the past as the grade when the child entered school and that there was no tracking in that grade. All subsequent learning is captured by later teacher effects and error terms.
\[
\begin{pmatrix}
A_{i(g-1)} \\
A_{i(g-2)} \\
e_{i(g-2)} \\
e_{i(g-1)}
\end{pmatrix}
\sim \mathcal{N}(\mu, \Sigma)
\]

In particular, the expected value of \(e_{ig}\) is a linear function of \(A_{i(g-1)}\). Given this, let \(er_{ig}\) be the residual from a regression of \(e_{ig}\) on \(A_{i(g-1)}\).

\[e_{ig} = \gamma A_{i(g-1)} + er_{ig}\]

where \(\gamma\) is the coefficient on lagged achievement.\(^{48}\)

Now we need to show that \(er_{ig}\) is uncorrelated with \(\tau_{ig}\), the current teacher. We start by assuming a specific functional form for \(\tau_{ig}\),

\[\tau_{1,i,g} = 1 \text{ if } A_{i(g-1)}>0 \text{ and } 0 \text{ otherwise.} \quad (B.2)\]

We then show that \(er_{ig}\) is uncorrelated with any function of \(A_{i(g-1)}\) and therefore is uncorrelated with \(\tau_{1,i,g}\).

By assumption, \(e_{ig}\) and \(A_{i(g-1)}\) are jointly normal. By construction \(er_{ig}\) is a linear function of these variables equal to \(e_{ig} - \lambda A_{ig}\). Thus \(er_{ig}\) and \(A_{i(g-1)}\) are also jointly normal. By construction, \(er_{ig}\) is also uncorrelated with \(A_{i(g-1)}\). Joint normality and zero correlation implies independence. This in turn means that \(er_{ig}\) is uncorrelated with any function of \(A_{i(g-1)}\) regardless of whether it is linear or nonlinear. Since, \(\tau_{ig}\), is a function of \(A_{i(g-1)}\), it is also uncorrelated with \(er_{ig}\).\(^{49}\)

Using the symbols from equation 2 in the main body of this paper (the omitted variable bias formula), this means that \(\text{cov}(e_{ig}\ast, \tau_{i(i,g)\ast}) = 0\). To see this note that \(er_{ig}\) is the residual that remains

\(^{48}\) We expect \(\gamma\) to be less than 0 since \(\text{cov}(e_{ig}, e_{i(g-1)})\) is less than 0.

\(^{49}\) In a more realistic scenario, \(\tau_{ig}\) would also depend on some additional variables. As long as they are also distributed independently of \(er_{ig}\), then \(er_{ig}\) will remain conditionally uncorrelated with \(\tau_{ig}\).
after regressing $e_{ig}$ on $A_{i(g-1)}$.\footnote{In the main body of this paper we discussed creating residuals by regressing each variable on the lagged test score and the other teacher dummies. In this model, there are no other teachers because there are only two teachers and one is omitted.} Thus, $e_{ig}$ is the same as $e_{ig*}$. Similarly, $\tau_{1i,g*}$ is the residual that remains after regressing $\tau_{1i,g}$ on $A_{i(g-1)}$. If $e_{ig*}$ is uncorrelated with $\tau_{1i,g}$ then it will also be uncorrelated with $\tau_{1i,g*}$ because $\tau_{1ig*}$ is a linear function of $\tau_{1ig}$ and $A_{i(g-1)}$ and $e_{ig*}$ is uncorrelated with both of those variables.

The negative correlation in errors means that students who scored lowest on the previous test will score somewhat higher than otherwise expected in the current period and vice versa. The coefficient estimate on $A_{i(g-1)}$ will be biased downwards, but, in this case, \textit{the negative correlation has no impact on the coefficient on $\tau_{1ig}$}.

Some readers might also be concerned about the plausibility of the data generation process we propose for tracking because it appears to depend on a linear function of lagged achievement. As noted above, tracking is necessarily a nonlinear function of lagged achievement. The functional form we propose allows for this. More precisely, one could think of tracking as a two-stage system in which tracking depends on a latent variable that, in turn, depends on lagged achievement. This can be written either by stage or in a single stage by substituting out for the latent variable (LV). Thus,

\begin{align*}
\text{Stage 1: } LV &= \beta T^* A_{i(g-1)} \\
\text{Stage 2: } \tau_{1ig} &= T(LV) \\
\text{Combined: } \tau_{1ig} &= T(\beta T^* A_{i(g-1)})
\end{align*}

where $\beta T^*$ is the coefficient on lagged achievement in Stage 1.

We have assumed that the first stage is linear. However, even if the first stage nonlinear this...
would not affect our argument because the second stage is nonlinear. Thus, nonlinearity in the first stage is not an issue. What is key to our argument—that a “plausible” data generation process can yield unbiased results—is that the relationship between the current error term and lagged achievement is linear.

**Non-Normally Distributed Lag Achievement Can Cause Bias**

The result above depends critically on grade 4 achievement being normally distributed. Without that property, bias can result. Grade 4 achievement is likely to be non-normally distributed if grade 4 teachers have impacts. As shown below, this implies that \( E(e_{ig-1}) \) is no longer a linear function of \( A_{i(g-1)} \), which in turn means that \( e_{i(g-1)} \) may depend on \( \tau_{i(g)} \) even after controlling for \( A_{i(g-1)} \). In addition, biased estimates may result, though the bias may be small.

To see why biased impacts might result, consider two sets of students: one set has the better lagged teacher in the previous period and the other has the worse lagged teacher. For each set of students, suppose that \( A_{i(g-1)} \) and \( e_{ig} \) were jointly normal. Then, for either group of students, the expected value of \( e_{ig} \) is a linear function of \( A_{i(g-1)} \). Let these functions be:

\[
E(e_{ig} | \tau_{i(g-1)} = 1, A_{i(g-1)}) = \lambda_1(A_{i(g-1)}) \quad \text{for students with the better lagged teacher in the previous period, and}
\]
\[
E(e_{ig} | \tau_{i(g-1)} = 0, A_{i(g-1)}) = \lambda_0(A_{i(g-1)}) \quad \text{for students with the omitted teacher.}
\]

Now consider the function describing the probability of having the better lagged teacher as a function of lagged achievement. This is necessarily a nonlinear function since \( \tau_{i(g-1)} \) is a discrete variable. Thus,

\[
\tau_{i(g-1)} = T_i(A_{i(g-1)})^{51}
\]

\[51\] This is more general than equation B.2.
Finally, the equation for the expected value of the error term as a function of $A_{i(g-1)}$ that combines both sets of students can be written as follows:

$$E(e_{ig} | A_{i(g-1)}) = \lambda_1(A_{i(g-1)})*T_1(A_{i(g-1)}) + \lambda_0(A_{i(g-1)})*T_0(A_{i(g-1)})$$

where $T_0(A_{i(g-1)}) = 1 - T_1(A_{i(g-1)})$.

Thus, $E(e_{ig} | A_{i(g-1)})$ is a nonlinear function of $A_{i(g-1)}$. Since both $E(e_{ig} | A_{i(g-1)})$ and $\tau_{ig}$ are nonlinear functions of $A_{i(g-1)}$, this suggests that they could be correlated even after controlling for $A_{i(g-1)}$ linearly. This could, in turn, cause bias in the estimated impact of teacher 1. Our simulation results support this point.
APPENDIX C

HOW LARGE IS THE BIAS CAUSED BY NEGATIVELY CORRELATED ERRORS?

In this appendix we present simulations to illustrate how large the bias might be with negatively correlated errors. We base our simulations on the Monte Carlo analyses presented in Appendix C of Rothstein (2010).\(^{52}\)

We assume that student achievement is generated by the following underlying model:

\[
AT_{ig} = \alpha + AT_{i(g-1)} + \lambda + \sum \tau_{i(g)} \beta + e_{ig} \quad \text{(C.1)}
\]

where \(AT_{ig}\) is true achievement for student \(i\) in grade \(g\), and \(\lambda\) is a school-by-grade effect for school \(s\).

Observed achievement equals true achievement plus random measurement error:

\[
AO_{ig} = AT_{ig} + e_{ig}
\]

with \(e_{ig}\) independent across grades and students.

To estimate the parameters in equation C.1, we use ordinary least squares and observed achievement. Thus,

\[
AO_{ig} = \alpha + AO_{i(g-1)} + \lambda + \sum \tau_{i(g)} \beta + ee_{ig}
\]

where \(ee_{ig}\) is the residual that is obtained when \(AT_{ig}\) and \(AT_{i(g-1)}\) are replaced by \(AO_{ig}\) and \(AO_{i(g-1)}\) from equation C.1.

\(^{52}\) Rothstein’s Monte Carlo simulation results are available in an online appendix (http://qje.oxfordjournals.org/content/suppl/2011/02/02/qjec.2010.125.1.175.DCSupplementaries/125-1-175-suppl.pdf) which shows how often his falsification test rejects VAM specifications with various parameter values and data generating processes. He finds that the VAM2 falsification test he developed tends not to reject the null hypothesis that future teachers have no effect on past grade test scores when students are randomly assigned to those future teachers. It does, however, tend to reject the null in cases where students are assigned to future teachers based in part on current grade achievement levels.
We generate data assuming $\lambda$ equals 1 but allow the estimated parameter on $A_{O, (g-1)}$ to differ from 1. This can occur because of measurement error in lagged achievement and/or if there is negative correlation in the true error terms ($\text{cov}(e_{ig}, e_{i(g-1)}) < 0$). In our simulations and those of Rothstein, true error terms and measurement errors are assumed to be normally distributed and uncorrelated with the current teacher effects. The measurement error terms are also uncorrelated with lagged achievement (observed and actual) while true errors are correlated with lagged achievement because they are correlated with each other across grades (as Rothstein assumes in his baseline model).

Like Rothstein, our simulation includes 400 schools, 4 teachers per grade per school and 20 students per class. This gives us a total sample of size of 1,600 teachers and 32,000 students in each replication. Because we include school fixed effects in our models, we can only estimate impacts for three of the four teachers in each school, and, thus, we estimate impacts for 1,200 teachers per replication. We use 1,000 replications.

Rothstein uses three rules for assigning students to current teachers. In the first, students are assigned randomly to teachers so, by definition, the assignment process generates no bias. He shows that in this case his falsification test accepts the null that future teachers have no effect on current teachers in the overwhelming majority of cases. In the second assignment rule, current students are tracked so that within a school a randomly chosen 50 percent of all students with

---

53 By allowing our estimate of $\lambda$ to differ from the value of one, we are using Rothstein’s “VAM2” model (Rothstein 2001). He also estimated another set of models (VAM1) where the outcome was the change in scores and the baseline score did not appear on the right hand side. That model is similar to the VAM2 model except that it constrains the coefficient estimate for $\lambda$ to be 1. Rothstein also estimated a third set of models (VAM3) with student fixed effects. We focus on the VAM2 specification to illustrate our findings.

54 Rothstein uses a test with a five percent significance level. This means that even when teachers are randomly assigned we would still expect that the Rothstein test would reject in five percent of trials due to random variation. Rothstein rejects more than five percent of the time with random assignment, perhaps because of the small sample of students. His results are shown in the first panel of his Appendix Table C.2.
each teacher are guaranteed to have the same prior teacher. The other 50 percent are randomly assigned to teachers. We call this “classroom tracking.” The third rule is used only for future teachers. Under this rule, students are tracked to future teachers based on current achievement and a random number. We refer to this as “test score tracking.” In our simulation, we use classroom tracking for the current teachers and test score tracking for future teachers.

Rothstein estimates a number of models that vary in terms of how students are tracked, the parameters used to generate the data, and the sample sizes. From these we pick his “baseline” parameters with tracking because they appear to represent a plausible set of parameters if one accepts the Rothstein hypothesis of negatively correlated errors.

Table C.1 summarizes the parameters we used in our simulation. The first two rows describe how students are tracked into classrooms in current and future grades. The remaining rows show the distribution of student-level error terms (the parts of achievement not explained by the model), distribution of the teacher effects, and other key parameters in the model.

55 Rothstein only uses “classroom tracking” for current teachers and not for future teachers. We follow this convention.

56 As noted earlier, Rothstein used the score with measurement error in his tracking models. We follow this convention as well in this appendix.

57 More precisely, each student’s percentile rank in the previous grade achievement distribution is computed. This rank is scaled from 0 to 1 and a random number with a uniform distribution between 0 and 1 is added. Students are sorted within school based on the sum of their percentile rank and random number as follows: students with the highest numbers are sorted into one teacher’s classroom until it is filled and then students with the next set of highest numbers are sorted into the next teacher’s classroom and so on until all the teachers’ classes are filled. When doing “test score” tracking for the current grade students, we maintain the Rothstein method of having classroom tracking for 50 percent of the students; the other 50 percent are tracked based on their test scores and a random number.
Table C.1. Student Assignment Methods and Data Parameters for Monte Carlo Simulations, Baseline Model of Rothstein

<table>
<thead>
<tr>
<th>Tracking</th>
<th>Classroom</th>
<th>Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>To grade g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>To grade g+1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Teacher Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{Var}(\beta_{tg}) ), ( \text{Var}(\beta_{t(g-1)}) )</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>( \text{Cov}(\beta_{tg},\beta_{t(g-1)}) )^a</td>
<td>-0.50</td>
<td></td>
</tr>
<tr>
<td><strong>School by Grade Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{Var}(\zeta_{g}) )</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>( \text{Var}(\zeta_{g(g-1)}) )^b</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>( \text{Cov}(\zeta_{tg},\zeta_{t(g-1)}) )</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td><strong>True Errors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{V}(e_{tg}) )</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>( \text{V}(e_{t(g-1)}) )^b</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>( \text{Cov}(e_{tg},e_{t(g-1)}) )</td>
<td>-0.25</td>
<td></td>
</tr>
<tr>
<td><strong>Measurement Error</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{V}(v_{tg}) ), ( \text{V}(v_{t(g)}) )</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>( \text{Cov}(v_{tg},v_{t(g-1)}) )</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Notes:

^a Correlations in teacher effects are based on teacher-level data. The correlations are between the current and past teachers that shared at least half of their students between grades. Since students are randomly assigned in the no-tracking model, the correlation is not used in that instance. Since we do not generate teacher effects for future teachers, the correlation is not used there either.

^b The variances of the true error and school/grade effects for the previous year may be larger than for the current year because prior year values capture impacts of all inputs for many years up to and including the prior year, whereas the current error/school-grade impacts only capture factors that affect student achievement over a one-year period starting at the end of the prior year and ending at the end of the current year.

In this appendix, we test the null hypothesis that \( \text{E}(\hat{\beta}_t) = \beta_t \) where \( \hat{\beta}_t \) is the estimate of \( \beta_t \), the coefficients for current teachers.\(^58\) We also present correlations between the true teacher

\(^58\) This is similar to the test used by Kane and Staiger (2008). However, their results may have been nonrepresentative if the teachers in their sample were primarily those for whom students were already being more or
effects and the estimated effects, with and without estimation error. More precisely, Table C.2 has the following information:

(a) Percentage of replications where F-tests reject the null that \( \text{E}(\hat{\beta}_t) = \beta_t \),\(^{59}\)

(b) Raw correlation between the estimated and true effects (correlations less than one could be due to bias or estimation error);

(c) Correlation between the estimated and true effects, after correcting for the estimation error (correlations of less than one imply bias).

Note that because we use a five percent significance rule, we would expect to reject the null in five percent of the replications using our test for bias.

(continued)

less randomly assigned. In contrast, we have simulated data where students are tracked to teachers based in part on prior achievement.

\(^{59}\) For each replication, we use an F-test to calculate the joint significance of the differences between estimated and true teacher effects. We include only three current teachers per school since we have a school dummy in the model. We used SAS software to do our simulations. Entering the true parameters directly into the hypothesis test commands in SAS was challenging given 1,200 teachers per simulation and 1,000 simulations. To get around this problem, we created data so that a regression using those data would produce the true parameters exactly as “estimates” with standard errors less than 1/30,000, as large as those estimated using the data described in Table D.1. Models were run combining the “parameter” data with the data described in Table D.1. This enabled us to run hypothesis tests comparing the true parameters (from the “parameter” data) to the estimated parameters.
Table C.2. Monte Carlo Analyses of Rothstein Falsification Test

<table>
<thead>
<tr>
<th>Percentage of replications where bias is found for current teachers ( \hat{\beta}_k = \beta_k )</th>
<th>6.6%**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted ( cor(\beta, \hat{\beta}) )</td>
<td>0.66</td>
</tr>
<tr>
<td>Adjusted ( cor(\beta, \hat{\beta}) )(^b)</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Notes:
** Indicates that results are statistically significant across the replications at the five percent level. To test the joint significance of the bias tests across replications, we used the Fisher chi-squared test (Fisher 1958).

\(^a\) The estimated current teacher effects are jointly statistically significant compared to 0 in 100 percent of the replications.

\(^b\) Adjusted \( cor(\beta, \hat{\beta}) \) is an estimate of the correlation that would be observed between \( \beta \) and \( \hat{\beta} \) in the absence of any estimation error (Spearman 1904; Goldhaber and Hansen 2010b).

The above table is based on a simulation with 400 schools, 4 classes per school, and 20 students per class in each replication. It uses 1,000 replications.

We would expect to find, and do find, evidence of a little bias based on the fact that the joint test across simulations is statistically significant at the five percent level. However, the percentage of simulations showing bias (i.e., the null is rejected) is very close to five percent, what we would get if the results were not biased. More importantly, the correlation between the estimated and true teacher effects is very close to 1 after adjusting for estimation error. We elaborate on why the bias can be small in Appendix B above.
APPENDIX D

MEASUREMENT ERROR CAN CAUSE NEGATIVE CORRELATION IN OBSERVED ERRORS

In this appendix we show that measurement error can cause observed error terms to be negatively correlated even if true errors are not correlated. We present this to illustrate how difficult it may be to differentiate between measurement error and negatively correlated errors in true achievement.

The effects of measurement error may differ depending on what assumptions are made about how students are tracked. To illustrate how measurement error might affect the results, we assume that teachers are selected based on true skills of students \( (A_{ig}) \), but that we do not observe those skills.\(^{60}\) Instead, we observe test score results that depend on true skills and, to some extent, on random factors that affect how well a student does on a given test independent of what they actually know on the day they are tested.\(^{61}\) This random factor—measurement error—is defined in equation (D.1) below.

\[
AO_{ig} = A_{ig} + v_g
\]

Now, if we estimate a VAM using \( AO_{ig} \) and \( AO_{i,(g-1)} \) in place of \( A_{ig} \) and \( A_{i,(g-1)} \), but do not adjust for measurement error in lagged achievement in order to obtain consistent estimates, the observed errors may be negatively correlated across grades even if the true errors are uncorrelated across grades. There is good reason to believe that this measurement error can

---

\(^{60}\) In Appendix C of his paper, Rothstein did simulations that incorporated measurement error but he assumed that teachers are selected based on the observed student test scores that include measurement error.

\(^{61}\) This can be thought of as the “luck of the draw” and might be related to the student simply getting unusually lucky when answering multiple choice questions or perhaps unlucky if, for example, they were distracted in an unusual way on the testing day or feeling a bit sick.
matter. For instance, adjustments for measurement error often increase the coefficient estimate on the lagged test score by 15 to 20 percent (Booker et al. 2008; Potamites et al. 2009).

As noted above, we do not observe current achievement, \( A_{ig} \). Rather, we observe current test scores, \( AO_{ig} \) which is equal to \( A_{ig} \) plus measurement error. Thus,

\[
AO_{ig} = \alpha + A_{i(g-1)} \lambda + \sum \tau_{i,g} \beta + e_{ig} + v_{ig}
\]

We also observe lagged test scores, \( AO_{i(g-1)} \) and not lagged achievement, \( A_{i(g-1)} \). Consequently, to see the equation, we can estimate we add and subtract \( AO_{i(g-1)} \) from a standard VAM equation:

\[
AO_{ig} = \alpha + A_{i(g-1)} \lambda + \sum \tau_{i,g} \beta + e_{ig} + v_{ig} + AO_{i(g-1)} \lambda - AO_{i(g-1)} \lambda
\]

Rearranging terms give us:

\[
AO_{ig} = \alpha + AO_{i(g-1)} \lambda + \sum \tau_{i,g} \beta + e_{ig} + v_{ig} + r_{ig}
\]

where

\[
r_{ig} = [A_{i(g-1)} - AO_{i(g-1)}] \lambda = -v_{i(g-1)} \lambda \quad \text{(D.2)}
\]

In other words, \( r_{ig} \) represents the error caused by using \( AO_{i(g-1)} \) in place of \( A_{i(g-1)} \).

Now the question is whether the combined error terms \( (e_{ig} + v_{ig} + r_{ig}) \) are negatively correlated across grades even if the originals \( (e_{ig} + v_{ig}) \) were not. The combined errors are negatively correlated because the first two terms are uncorrelated over time (by assumption), while \( r_{ig} \) is negatively correlated with the measurement error from the previous period \( (v_{i(g-1)}) \), as shown by equation D.2.\(^62\)

\(^62\) This finding also suggests that measurement error may help to explain why both Rothstein and Kodel and Betts find no evidence of student fixed effects, which would cause error terms to be more positively correlated across grades. However, as discussed here, measurement error might cause observed errors to be more negatively correlated than the true errors. The net effect might be no evidence of student fixed effects even if they were there in reality.
The negatively correlated errors shown by equation D.2 mean that measurement error can cause the estimated correlation in residuals (observed errors) to be negative even if the true errors are not negatively correlated.\textsuperscript{63}

\textsuperscript{63} The observed errors will differ from those discussed here because the coefficient estimates will be estimated with error and perhaps bias, depending on the estimation method used.